# 谢友柏设计科学研究基金项目
# 年度报告

项目名称：互联网群智协同创新环境下有效用户知识贡献行为演化及其驱动模式研究

负 责 人：梁若愚

依托单位：江南大学

通讯地址：江苏省无锡市滨湖区蠡湖大道 1800 号江南大学设计学院

邮政编码：214122

电子邮件：lryasa@tju.edu.cn

电　　话：17315530191

报送日期：2021.4

1. 年度计划研究内容

2020.1—2020.4　全面搜集相关文献，归纳研究目标社区功能结构、内容产生形式及数据类型，完善技术路线和实施方案。

2020.5—2020.12　采集目标社区典型创新项目用户相关数据，统计分析用户贡献行为，构建行为演化总体模型；构建创新词库，研究有效参与者评价指标体系构建方法。

2. 年度研究进展及成果

2.1 年度研究进展

根据研究计划，本年度主要围绕群智协同创新社区的功能特征、版块架构、内容产生形式、数据类型、用户构成与行为特点、有效用户评价指标体系与识别方法等展开分析与探索。

2.1.1 群智协同创新社区结构与用户分析

通过对 20 余个典型群智协同创新平台的分析与研究，归纳出了目前应用较多的 3 类群智创新平台运营模式：

（1）商铺型

该类型平台以猪八戒网、Mechanical Turk、designhill 等为代表，用户可通过 web 页面与移动终端 APP 访问网站，其设计服务可包括品牌策划、图形设计、产品研发、用户研究等。该类平台通过交易形式为需求方提供创新服务，常见的交易方式可包括 1 对 1 服务、购买封装好的设计方案、招投标等。该类平台一般包含两种参与者，即需求方与资源提供方，两者在交易过程中构建了一种雇佣关系。平台运营方为交易双方提供协作环境与保障，通过抽成、会员年费、广告费

等方式获取收益。为了保证平台服务的质量，运营方会对拟加入平台的资源提供方进行审核，具备相应能力的个人/团队方可入驻平台。

（2）专业社区型

该类平台以 LocalMotors、OpenIDEO、Innocentive、P&G connect+develop 等为代表，用户可通过 web 页面访问网站，其服务类型涵盖面极广，可包括产品研发、设计策划、公共管理、组织决策、信息咨询等各方面社会民生主题。该类平台包含三种参与者，即需求方、资源提供方、行业专家。平台既提供盈利性服务，也提供公益性服务，常见的组织流程如下：需求方可通过在社区发帖、发布博客的方式在平台提出问题，有解决能力的资源提供方可通过回帖、发起主题讨论等渠道帮助需求方解决问题；平台运营方定期分析网站产生的数据流，从中选取具有一定社会、经济价值的问题作为官方主题发布在平台，并通过竞赛、招标等形式招募资源提供方协作完成相关任务，在此过程中，驻站的行业专家可以为参与者提供必要的支持。平台方可通过收益分成、广告费、服务费等方式获取收益。运营方会对行业专家的资质、需求方提出的问题内容等进行审核，以保证平台的服务质量与稳定运行。

（3）大众社区型

该类平台以小米社区、Dell Ideastorm、Ducati 社区等为代表，其本质上是企业为实现产品推广、用户交流、经验分享、口碑营销等目标而构建的品牌社区，用户可通过 web 页面访问网站。在设计与创新方面，该类社区的作用主要是通过一定的激励手段，促进用户贡献与产品开发相关的内容，如消费者对产品的评价、期望、不满等，用户在产品使用过程中的体验以及所发现的不足与缺陷，用户所掌握的知识、技能、问题解决方案等。该类平台的参与人员相对复杂，包括

消费者、品牌追随者、经销商、技术爱好者等。平台用户可以在社区中自由发帖，阐述个人观点、诉求等，企业与其他用户可对帖子进行回应。企业拥有对社区的完整管理权限，可通过推广主题、促进研讨、信息匹配与推送、奖励以及筛查发帖内容、管理成员行为权限、定向回应等手段影响社区的话题方向，使平台完全为企业的经济目标服务。

2.1.2 群智协同创新环境下的众包设计模式与定义

（1）众包设计的定义

群智协同创新的主要实施方式是通过众包策略，招募具有相应能力的个体与团体协同解决设计与创新问题。根据理论分析与案例研究，可发现目前众包设计主要可包含两个亚类，即面向制造业龙头企业前沿技术众包设计创新模式与基于第三方平台的个性化众包设计能力拓展模式。

前者是将设计研发过程中出现的问题通过内部平台发布以招募内外部资源协同解决的做法，需要综合考虑设计问题的类型、信息传播方式、协同方式、外部资源的保密性等因素；后者是将设计工作通过互联网平台部分或全部的分包给具备相应能力、资源的组织或个体的做法，需要综合考虑总体设计工作的规模、任务分解方式、任务粒度、能力需求、任务与能力匹配、参与人群类别/作用、个体能力识别、任务推送、参与者管理与激励、任务发包方式、验收标准、验收方式、成本等因素，全面优化设计质量（Q）、效率（T）、经济性（C）、设计资源利用率（R）、产品性能（F）等目标函数，求得其最佳平衡点。

（2）众包设计的主要目标：

1）精准应对用户个性化需求，优化产品定制属性。随着经济社会的快速发展，人们对于各类产品在功能、外观等方面的个性化需求也越来越突出，在这样

一种背景下，传统的组织内部封闭式创新已经很难适应用户日新月异的消费诉求。众包模式可以有效地打破组织边界，获取广泛的创新思路，且大多数设计参与者都拥有较为丰富的产品使用经验与前卫的使用需求，在一定程度上能够代表市场的发展趋势。因而，众包设计模式在有效匹配用户需求方面相较传统设计模式具有明显优势。

2）激活社会性设计资源，提升设计创新效率。传统设计模式对于专业设计资源如设计师、工程师等依赖性较强，实施门槛较高。众包模式能够将设计任务粒度进一步缩小，从而降低设计活动入门级别，促进广域资源参与其中，实现设计任务的并行化，缩短任务流程，提升实施效率。

3）获取更广泛创意资源，拓宽设计问题解决途径。传统设计模式在创意、思路等方面往往受限于设计人员的知识、经验、能力、天赋等因素，问题解决途径相对有限。众包模式可以吸引更为广泛的大众群体加入设计活动，从而显著扩增创意、方案数量规模，为设计问题提供更多、更优质的解决渠道。

（3）众包设计的主要内容

众包设计实际上是面向产品、需求方、使用方等多环节、多方面的系统化设计模式。图 1 为众包设计模式所面向的全过程，各个环节有机结合形成一个完整的生态体系。

1）设计任务规划与分解

需求方提出设计任务后，首先需要对项目实施流程进行总体规划，如参与人员与成本规模、项目周期、监督管理与支持方式等，设置专门的项目负责人员，形成规范化的管理文件。第二，根据产品设计工作的复杂程度、专业水平、知识领域、潜在用户、能力需求等设定计划招募的参与人员类型如设计师、工程师、

高校师生、资深用户等，并定位目标人群的网络聚集地如各类专业论坛、开放式众包平台、社交网站等。第三，根据设计任务的特点选择适宜的任务分解理论与工具，规划任务分解的方式与步骤。第四，结合任务分解方式、参与人群特征及能力分布等分析任务粒度，将总体任务拆解成适合个体或小型团队完成的子任务。第五，分析每个子任务所需要的能力种类与水平层次，形成标准化指导文本。

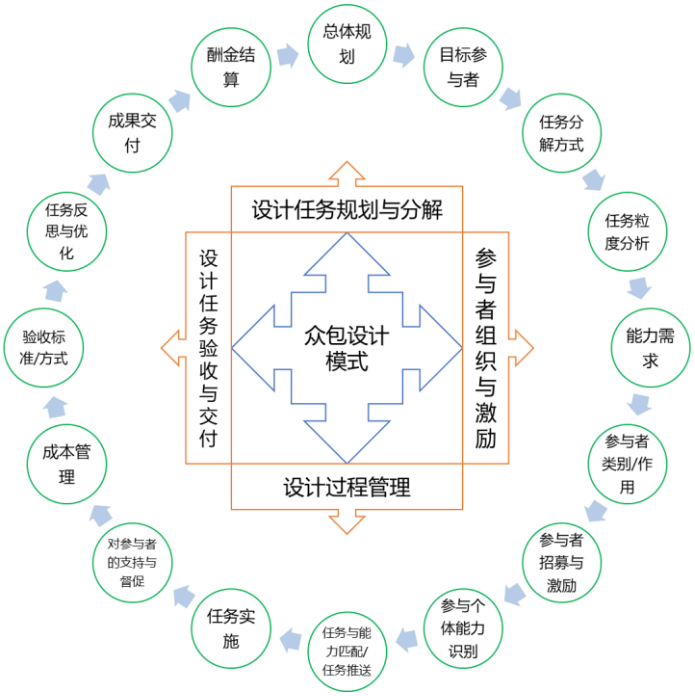



图 1 众包设计模式实施全过程

2）参与者组织与激励

在任务规划与分解的同时，需要开展参与者的招募与组织工作：第一，通过数据挖掘技术采集目标人群网络聚集地的用户留存数据、注册信息等，结合调查访谈等手段确定目标参与人群，构建有效参与者资源池，根据专长、能力等对潜在参与者进行分类，分析各类人员在设计项目中可能发挥的作用。第二，综合运用物质与精神激励手段，招募资源池中的有效参与者，与之形成长久而稳定的联系纽带，将个体粘性维持在较高水平。第三，基于用户贡献内容、注册信息、关

注版块等，分析个体的能力域，通过调查访谈等方法进一步确定参与者的能力专长。第四，根据子任务的能力需求，向处于对应能力域的个体及小型团队推送相关任务邀请。

3）设计过程管理

完成任务分解与人员招募工作后，即进入项目实施阶段：第一，根据管理文件制定各子任务进程节点、沟通机制、指导方式（如参与式指导、周期性指导）等。第二，按照规则对项目实施情况进行监督，为参与者提供技术、工具、方法、理论等方面的指导与培训。第三，严格遵照预算计划对研发成本进行管理，单线程项目应动态优化投入方式；多线程项目应适时淘汰进度缓慢、成果不理想的项目组。第四，通过酬金、优惠权限、奖品等物质激励手段，保持参与者/团队的积极性。

4）设计任务验收与交付

动态评价各子任务的实施效果，及时验收已经完成的细分任务：第一，根据管理文件对子任务进行评估与验收，必要情况下可适当调整验收规则。第二，根据评估结果，为参与者提供成果反馈，对于无法满足验收条件的子任务要及时介入指导、优化乃至更换承担团队。第三，通过评估后即根据预先约定交付相关成果，明确产权归属，签订产权相关法律文件。第四，根据管理文件支付酬金或项目尾款。

2.1.3 有效用户评价指标体系与识别方法

在专业社区与大众社区型群智协同创新平台环境下，用户可以自由注册成为平台的提供方。然而，不同用户在能力水平与知识储备等方面都存在巨大的差异性，且相当一部分平台用户日常贡献的有效内容（包括与设计创新相关的技术、

方法、方案、知识、观点、需求等）较少，其产生的主要内容为社交、娱乐等无效信息，这类个体在平台中属于价值、效用较低的用户，需求方在选择资源提供方时应当尽量规避此类用户。

针对这一问题，本研究构建了用于评价社区用户有效性的指标体系，并提出了一种基于变异系数法、层次分析法与灰色关联度分析的有效用户识别方法。其具体步骤如下：

（1）通过数据挖掘技术获取用户在社区中的留存数据，包括行为统计数据与贡献内容等；

（2）将用户积分、等级以及用户贡献内容中的产品特征词汇、产品问题词汇、情绪词汇、评论有效长度、评论时效性、用户能力评估等作为评价用户有效性的指标，构建评价指标体系；

（3）将各类指标值进行归一化处理，使用变异系数法/层次分析法确定各指标的权重值，计算各样本用户的指标值；

(4) 使用灰色关联度分析计算候选用户指标序列与参考序列之间的关联度，关联度值高于预先设置好的阈值的就是群智协同创新有效参与者。

2.2 年度成果

本年度形成的主要成果包括发明专利一部及论文一篇，其中专利目前处于公布及进入实质审查阶段（发明名称：一种识别互联网技术社区中众包设计有效参与者的方法，申请号：202011531547.6，实审发文序号：2021041400755460），论文（题目：How to find the key participants in crowdsourcing design? Identifying lead users in the online context using user-contributed content and online behavior analysis，提出了一种面向群智协同创新社区的有效用户识别方法，并

通过实证检验了所提出方法的有效性）已投往 SSCI 期刊 Technology Analysis & Strategic Management，目前处于一审阶段。附件为专利稿（包括受理函与实审通知书）与论文初稿，由于相关成果尚未发表，请暂缓公开年度报告。

（成果可以是未发表的研究报告、论文稿、专利稿，成果发表均需标注有“谢友柏设计科学研究基金资助”字样）

附：上述研究成果全文或实体，成果不能在互联网上传递的可以邮寄到学委会秘书组。

# 国 家 知 识 产 权 局

**215000**

中国（江苏）自由贸易实验区苏州片区苏州工业园区崇文路199号富华大厦501室

苏州市中南伟业知识产权代理事务所（普通合伙） 郭磊

(0512-65882644)

发文日：

2020年12月23日

申请号或专利号：**202011531547.6**　　发文序号：**2020122300731480**

## 专 利 申 请 受 理 通 知 书

根据专利法第28条及其实施细则第38条、第39条的规定，申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日、申请人和发明创造名称通知如下：

申请号：202011531547.6

申请日：2020年12月22日

申请人：江南大学

发明创造名称：一种识别互联网技术社区中众包设计有效参与者的方法

经核实，国家知识产权局确认收到文件如下：

实质审查请求书 每份页数:1页 文件份数:1份

说明书摘要 每份页数:1页 文件份数:1份

发明专利请求书 每份页数:4页 文件份数:1份

说明书附图 每份页数:1页 文件份数:1份

说明书 每份页数:6页 文件份数:1份

权利要求书 每份页数:2页 文件份数:1份 权利要求项数： 7项

提示：

1. 申请人收到专利申请受理通知书之后，认为其记载的内容与申请人所提交的相应内容不一致时，可以向国家知识产权局请求更正。

2. 申请人收到专利申请受理通知书之后，再向国家知识产权局办理各种手续时，均应当准确、清晰地写明申请号。

3. 国家知识产权局收到向外国申请专利保密审查请求书后，依据专利法实施细则第9条予以审查。

审 查 员：自动受理　　审查部门：专利局初审及流程管理部

# 国 家 知 识 产 权 局


**215000**

中国（江苏）自由贸易实验区苏州片区苏州工业园区崇文路 199 号富华大厦 501 室 苏州市中南伟业知识产权代理事务所（普通合伙）

郭磊(0512-65882644)


发文日：

2021 年 04 月 19 日

申请号或专利号：**202011531547.6**　　发文序号：**2021041400755460**

申请人或专利权人： 江南大学

发明创造名称： 一种识别互联网技术社区中众包设计有效参与者的方法

## 发 明 专 利 申 请 公 布 及 进 入 实 质 审 查 阶 段 通 知 书

上述专利申请，经初步审查，符合专利法实施细则第 44 条的规定。根据专利法第 34 条的规定，该申请在 37 卷 1601 期 2021 年 04 月 13 日专利公报上予以公布。

根据申请人提出的实质审查请求，经审查，符合专利法第 35 条及实施细则第 96 条的规定，该专利申请进入实质审查阶段。

提示：

1. 根据专利法实施细则第 51 条第 1 款的规定，发明专利申请人自收到本通知书之日起 3 个月内，可以对发明专利申请主动提出修改。

2. 申请人可以访问国家知识产权局政府网站（www.cnipa.gov.cn）,在专利检索栏目中查询公布文本。如果申请人需要纸件申请公布单行本的纸件，可向国家知识产权局请求获取。

3. 申请文件修改格式要求：

对权利要求修改的应当提交相应的权利要求替换项，涉及权利要求引用关系时，则需要将相应权项一起替换补正。如果申请人需要删除部分权项，申请人应该提交整理后连续编号的部分权利要求书。

对说明书修改的应当提交相应的说明书替换段，不得增加和删除段号，仅只能对有修改部分段进行整段替换。如果要增加内容，则只能增加在某一段中；如果需要删除一个整段内容，应该保留该段号，并在此段号后注明：“此段删除”字样。段号以国家知识产权局回传的或公布/授权公告的说明书段号为准。

对说明书附图、摘要、摘要附图修改的应当提交相应的说明书附图、摘要、摘要附图替换页。

同时，申请人应当在补正书或意见陈述书中标明修改涉及的权项、段号、页。

审　查　员：自动审查　　审查部门：专利局初审及流程管理部

联系电话：010-62084704



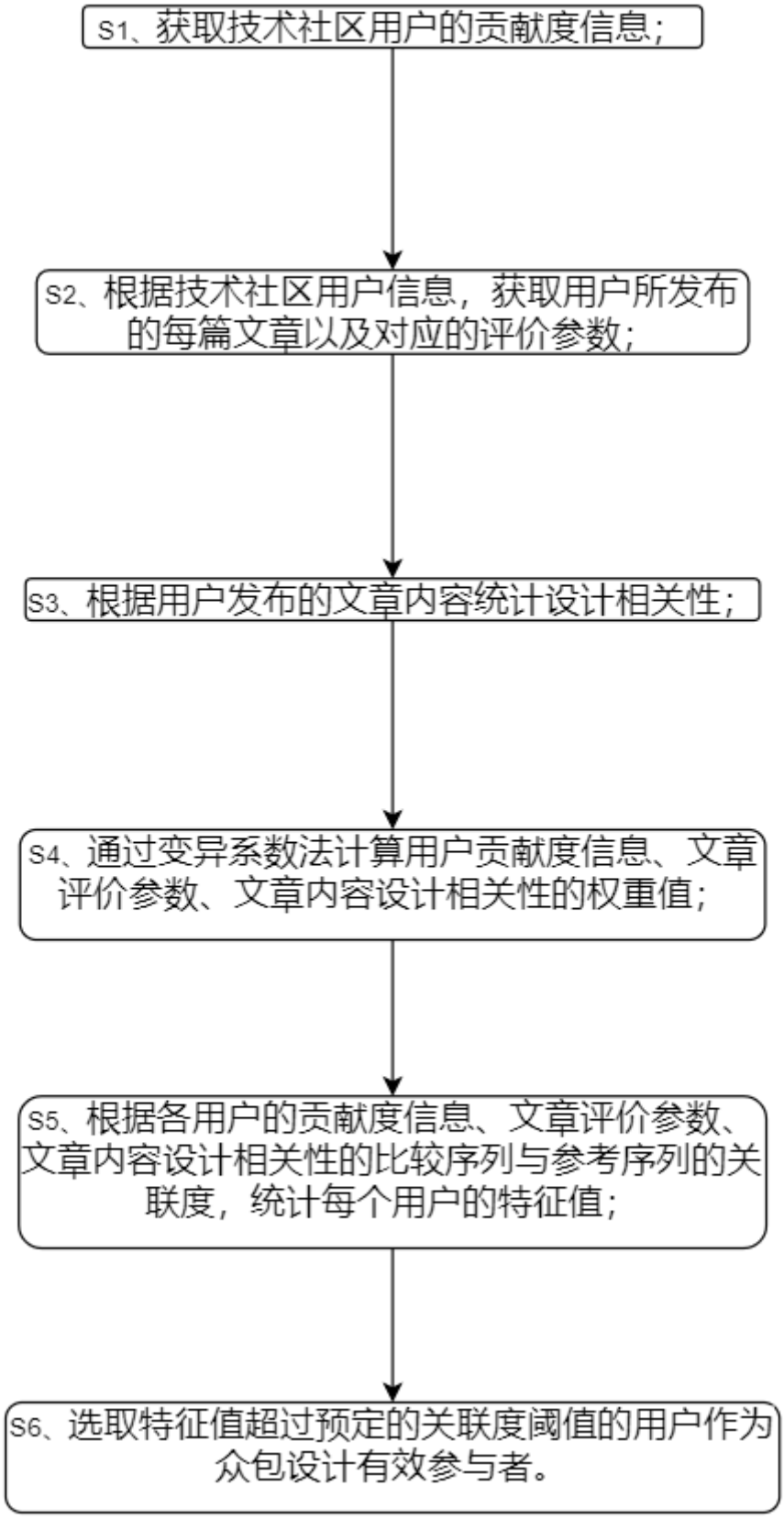
纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路 6 号　国家知识产权局专利局受理处收

电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

# 说　明　书　摘　要

本发明涉及互联网技术的技术领域，特别是涉及一种识别互联网技术社区中众包设计有效参与者的方法，其能够剔除无用信息，准确甄别具备设计创新能力的有效用户，招募高质量设计参与者，同时具有完善的对于众包设计有效参与者的评价指标体系；包括以下步骤：S1、获取技术社区用户的贡献度信息；S2、根据技术社区用户信息，获取用户所发布的每篇文章以及对应的评价参数；S3、根据用户发布的文章内容统计设计相关性；S4、通过变异系数法计算用户贡献度信息、文章评价参数、文章内容设计相关性的权重值；S5、根据各用户的贡献度信息、文章评价参数、文章内容设计相关性的比较序列与参考序列的关联度，统计每个用户的特征值。

# 摘　要　附　图


S1、获取技术社区用户的贡献度信息；
S2、根据技术社区用户信息，获取用户所发布的每篇文章以及对应的评价参数；
S3、根据用户发布的文章内容统计设计相关性；
S4、通过变异系数法计算用户贡献度信息、文章评价参数、文章内容设计相关性的权重值；
S5、根据各用户的贡献度信息、文章评价参数、文章内容设计相关性的比较序列与参考序列的关联度，统计每个用户的特征值；
S6、选取特征值超过预定的关联度阈值的用户作为众包设计有效参与者。

# 权　利　要　求　书

1、一种识别互联网技术社区中众包设计有效参与者的方法，其特征在于，包括以下步骤：

S1、获取技术社区用户的贡献度信息；

S2、根据技术社区用户信息，获取用户所发布的每篇文章以及对应的评价参数；

S3、根据用户发布的文章内容统计设计相关性；

S4、通过变异系数法计算用户贡献度信息、文章评价参数、文章内容设计相关性的权重值；

S5、根据各用户的贡献度信息、文章评价参数、文章内容设计相关性的比较序列与参考序列的关联度，统计每个用户的特征值；

S6、选取特征值超过预定的关联度阈值的用户作为众包设计有效参与者。

2、如权利要求 1 所述的一种识别互联网技术社区中众包设计有效参与者的方法，其特征在于，所述 S4 中将“用户积分”、“帖子转发数”、“帖子评论数”、“用户贡献的文本内容中的产品结构/功能/外观相关词汇”、“技术相关词汇”、“设计相关词汇”、“贡献内容的有效长度”、“贡献内容时效性”作为评价参与者有效性的指标，使用变异系数法计算每个评价指标的重要性权重。

3、如权利要求 2 所述的一种识别互联网技术社区中众包设计有效参与者的方法，其特征在于，所述各项指标的统计计算方法如下：

“用户积分”、“帖子转发数”、“帖子评论数”为直接取自互联网技术社区的统计数据；

“用户贡献的文本内容中的产品结构/功能/外观相关词汇”、“技术相关词汇”、“设计相关词汇”为用户发布的文章中出现对应词汇的数量；

“贡献内容的有效长度”：$R=\dfrac{\lg(N_b+N_c+N_d)}{\lg N_a}$；

其中 $N_a$、$N_b$、$N_c$、$N_d$分别代表用户发布的文章中的词汇数、产品结构/功能/外观相关词汇数、技术相关词汇数、设计相关词汇数；

“贡献内容时效性”：近两个月内发布的文章记为 2，两个月以前发布的记为 1。

4、如权利要求 3 所述的一种识别互联网技术社区中众包设计有效参与者的方法，其特征在于，各项指标的变异系数公式为：

$$V_i = \frac{\sigma_i}{\bar{x}_i}$$

其中 i=1,2,…,n ；$V_i$是第 $i$ 项指标的异变系数，即标准差系数；$\sigma_i$是第 $i$ 项指标的标准差；$\bar{x}_i$是第 $i$ 项指标的平均数。

5、如权利要求 4 所述的一种识别互联网技术社区中众包设计有效参与者的方法，其特征在于，各项指标的权重计算公式为：

$$W_i = \frac{V_i}{\sum_{i=1}^{n} V_i}$$ 。

6、如权利要求 5 所述的一种识别互联网技术社区中众包设计有效参与者的方法，其特征在于，选取研究样本中各个指标的最优值组成参考序列，利用灰色关联分析计算候选用户指标序列与参考序列之间的关联度，关联度值高于预先设置好的阈值的就是众包设计有效参与者。

7、如权利要求 1 所述的一种识别互联网技术社区中众包设计有效参与者的方法，其特征在于，所述阈值可以根据众包设计的要求进行修改。

# 说　　明　　书

## 一种识别互联网技术社区中众包设计有效参与者的方法

### 技术领域

本发明涉及互联网技术的技术领域，特别是涉及一种识别互联网技术社区中众包设计有效参与者的方法。

### 背景技术

在线众包是 2006 年前后兴起的一种新商业模式，其目标是将复杂工作拆解后通过互联网分配给大众以协同方式完成来创造经济价值。众包设计模式则是将设计工作通过互联网平台部分或全部的分包给具备相应能力、资源的组织或个体的做法，需要综合考虑总体设计工作的规模、任务分解方式、能力需求、参与人群类别/作用、参与者管理与激励、任务发包方式、验收标准/方式、成本等因素。众包设计得以实施的前提是定位并招募到具备设计相关能力的用户个体，这些个体的互联网聚集地是各类技术社区如站酷、CSDN 论坛、中国机械社区、UI 中国等。

近年来，越来越多的企业开始借助众包设计模式开展产品研发活动，以应对逐渐多元化的市场需求发展趋势，然而，该模式在具体实施过程中，发明人发现现有技术中至少存在如下问题：

（1）由企业/第三方运营的专业众包设计服务平台往往知名度较小、公众影响力弱，很难招募到足质足量的设计参与者；

（2）互联网技术社区中包含大量无效用户如广告发布者、灌水者等，众包设计活动组织者缺乏能够准确甄别具备设计创新能力的有效用户的策略与方法；

（3）缺乏对于众包设计有效参与者的评价指标体系。

### 发明内容

为解决上述技术问题，本发明提供一种能够剔除无用信息，准确甄别具备设计创新能力的有效用户，招募高质量设计参与者，同时具有完善的对于众包设计有效参与者的评价指标体系的识别互联网技术社区中众包设计有效参与者的方法。

本发明的一种识别互联网技术社区中众包设计有效参与者的方法，包括以下步骤：

S1、获取技术社区用户的贡献度信息；

S2、根据技术社区用户信息，获取用户所发布的每篇文章以及对应的评价参数；

S3、根据用户发布的文章内容统计设计相关性；

S4、通过变异系数法计算用户贡献度信息、文章评价参数、文章内容设计相关性的权重值；

S5、根据各用户的贡献度信息、文章评价参数、文章内容设计相关性的比较序列与参考序列的关联度，统计每个用户的特征值；

S6、选取特征值超过预定的关联度阈值的用户作为众包设计有效参与者。

本发明的一种识别互联网技术社区中众包设计有效参与者的方法，所述 S4 中将“用户积分”、“帖子转发数”、“帖子评论数”、“用户贡献的文本内容中的产品结构/功能/外观相关词汇”、“技术相关词汇”、“设计相关词汇”、“贡献内容的有效长度”、“贡献内容时效性”作为评价参与者有效性的指标，使用变异系数法计算每个评价指标的重要性权重。

本发明的一种识别互联网技术社区中众包设计有效参与者的方法，所述各项指标的统计计算方法如下：

“用户积分”、“帖子转发数”、“帖子评论数”为直接取自互联网技术社区的统计数据；

“用户贡献的文本内容中的产品结构/功能/外观相关词汇”、“技

术相关词汇”、“设计相关词汇”为用户发布的文章中出现对应词汇的数量；

“贡献内容的有效长度”：$R=\frac{\lg(N_b+N_c+N_d)}{\lg N_a}$；

其中 $N_a$、$N_b$、$N_c$、$N_d$ 分别代表用户发布的文章中的词汇数、产品结构/功能/外观相关词汇数、技术相关词汇数、设计相关词汇数；

“贡献内容时效性”：近两个月内发布的文章记为 2，两个月以前发布的记为 1。

本发明的一种识别互联网技术社区中众包设计有效参与者的方法，各项指标的变异系数公式为：

$$V_i=\frac{\sigma_i}{\bar{x}_i}$$

其中 i=1,2,…,n ；$V_i$是第 $i$ 项指标的异变系数，即标准差系数；$\sigma_i$是第 $i$ 项指标的标准差；$\bar{x}_i$是第 $i$ 项指标的平均数。

本发明的一种识别互联网技术社区中众包设计有效参与者的方法，各项指标的权重计算公式为：

$$W_i=\frac{V_i}{\sum_{i=1}^{n}V_i}。$$

本发明的一种识别互联网技术社区中众包设计有效参与者的方法，选取研究样本中各个指标的最优值组成参考序列，利用灰色关联分析计算候选用户指标序列与参考序列之间的关联度，关联度值高于预先设置好的阈值的就是众包设计有效参与者。

本发明的一种识别互联网技术社区中众包设计有效参与者的方法，所述阈值可以根据众包设计的要求进行修改。

与现有技术相比本发明的有益效果为：能够剔除无用信息，准确甄别具备设计创新能力的有效用户，招募高质量设计参与者，同时具有完善的对于众包设计有效参与者的评价指标体系。

**附图说明**

图1是本发明的逻辑流程图。

**具体实施方式**

下面结合附图和实施例，对本发明的具体实施方式作进一步详细描述。以下实施例用于说明本发明，但不用来限制本发明的范围。

通过数据挖掘技术（发明不中不体现）获取技术社区用户的等级信息，包括积分、贡献值、级别等，以及贡献内容如发帖、回帖等文本，以及这些帖子下面的评论数、转发数等；将“用户积分”、“帖子转发数”、“帖子评论数”、“用户贡献的文本内容中的产品结构/功能/外观相关词汇”、“技术相关词汇”、“设计相关词汇”、“贡献内容的有效长度”、“贡献内容时效性”作为评价参与者有效性的指标，其中各项指标的统计计算方法如下：

“用户积分”、“帖子转发数”、“帖子评论数”为直接取自互联网技术社区的统计数据；

“用户贡献的文本内容中的产品结构/功能/外观相关词汇”、“技术相关词汇”、“设计相关词汇”为用户发布的文章中出现对应词汇的数量；

“贡献内容的有效长度”：$R=\dfrac{\lg(N_b+N_c+N_d)}{\lg N_a}$；

其中 $N_a$、$N_b$、$N_c$、$N_d$ 分别代表用户发布的文章中的词汇数、产品结构/功能/外观相关词汇数、技术相关词汇数、设计相关词汇数；

“贡献内容时效性”：近两个月内发布的文章记为 2，两个月以前发布的记为 1。

使用变异系数法计算每个评价指标的重要性权重，步骤如下：

1）、各项指标的变异系数公式：

$$V_i=\frac{\sigma_i}{\bar{x}_i}\quad(i=1,2,\cdots,n)$$

式中$V_i$是第 $i$ 项指标的异变系数，也被称为标准差系数；$\sigma_i$是第 $i$ 项指标的标准差；$\bar{x}_i$是第 $i$ 项指标的平均数。

2）、各项指标的权重为：

$$W_i = \frac{V_i}{\sum_{i=1}^{n} V_i};$$

选取研究样本中各个指标的最优值组成参考序列；利用灰色关联分析计算候选用户指标序列与参考序列之间的关联度，灰色关联分析计算具体如下：

步骤 1、原始数据无量纲化处理：

$X_k(i)$ 与 $Y(i)$ 分别表示比较序列与参考序列，$k=1,2,\cdots,m$，$i=1,2,\cdots,n$，$m$ 表示待排序用户数量，$n$ 表示指标个数；使用均值化方法对原始数据进行规范化处理，公式如下：

$$x'_{ki} = \frac{x_{ki}}{x_i} \quad k=1,2,\ldots,m,\ i=1,2,\ldots,n$$

$$y'_i = \frac{y_i}{x_i} \quad i=1,2,\ldots,n$$

式中：$x_{ki}$ 为第 $k$ 个目标用户的第 $i$ 个指标的量化值；$x_i$ 为待排序用户集当中第 $i$ 个指标的平均值；$y_i$ 为参考序列当中第 $i$ 个指标的量化值；需要注意，使用均值法进行规范化处理的数据，在后续计算过程中需要将某些指标值与参考序列中对应指标值相同的用户剔除掉，以免产生无效的结果；

步骤 2、计算灰色关联系数；

灰色关联系数的计算公式如下所示：

$$r_{ki} = \frac{\Delta 1 + \zeta \Delta 2}{\Delta 3 + \zeta \Delta 2} \quad k=1,2,\ldots,m,\ \ i=1,2,\ldots,n$$

式中：$\Delta 1 = \min_{k\in M}\left\{\min_{i\in N}\left|Y'(i) - X'_k(i)\right|\right\}$，$\Delta 2 = \min_{k\in M}\left\{\max_{i\in N}\left|Y'(i) - X'_k(i)\right|\right\}$，$\Delta 3 = \left|Y'(i) - X'_k(i)\right|$，其中 $M = \{1,2,\ldots,m\}$，$N = \{1,2,\ldots,n\}$；$\zeta$为分辨系数，在[0,1]区间取值，通常情况下取$\zeta$=0.5；

步骤 3 计算灰色关联度；

采用加权的方法计算灰色关联度，公式如下所示：

$$d_k = \sum_{i=1}^{n} w_i r_{ki} \qquad k \in M$$

式中：$w_i$为指标权重值；根据各个用户关联度的大小进行排序，得到用户有效性排序；关联度值高于预先设置好的阈值的就是众包设计有效参与者。

以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明技术原理的前提下，还可以做出若干改进和变型，这些改进和变型也应视为本发明的保护范围。

# 说　明　书　附　图

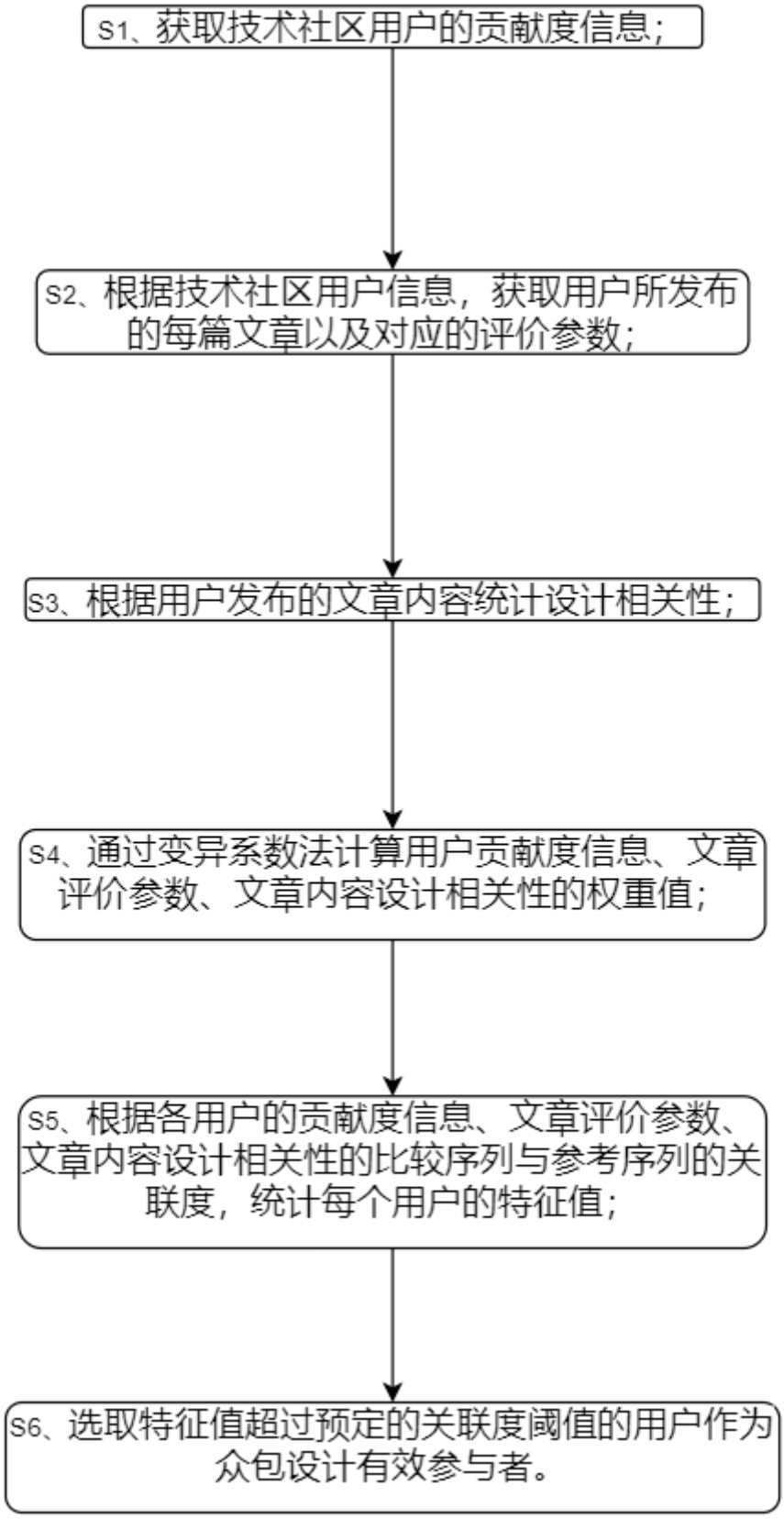

S1、获取技术社区用户的贡献度信息；
S2、根据技术社区用户信息，获取用户所发布的每篇文章以及对应的评价参数；
S3、根据用户发布的文章内容统计设计相关性；
S4、通过变异系数法计算用户贡献度信息、文章评价参数、文章内容设计相关性的权重值；
S5、根据各用户的贡献度信息、文章评价参数、文章内容设计相关性的比较序列与参考序列的关联度，统计每个用户的特征值；
S6、选取特征值超过预定的关联度阈值的用户作为众包设计有效参与者。

图 1

**Main Document − Anonymous**

# How to find the key participants in crowdsourcing design? Identifying lead users in the online context using user-contributed content and online behavior analysis

**Abstract:** Lead users are the most valuable innovation sources in crowdsourcing design; how to identify these users is a research hotspot in the field of design and management. Existing approaches to discover lead users in the context of online community, such as manual method and ordering algorithm, have some limitations, for instance, low coverage and accuracy. To address these deficiencies, this paper proposes a method which applying text-mining techniques, analysis of user's behavior and contributed-content to identify lead users. We suggest a three-step analytical approach: First, a criterion system to evaluate the user's leading-edge status is constructed. Second, we utilize a fuzzy analytical hierarchy process to assess the weighted value of each indicator and develop the reference sequence of the indicators. Third, grey relational analysis is employed to analyze the correlations between users' indicators and reference sequences, and lead users are recognized based on the ranking of each user's correlation. An empirical analysis is used to examine the effectiveness of the proposed method. The results reveal that the method has good precision and recall rate, it can automatically process large-scale data and has no strict requirements for respondents. Finally, the paper discusses the limitations and provides possible directions for future research.

**Keywords:** crowdsourcing design community; identification method of lead users; behavior analysis; analysis of contribution content; fuzzy analytical hierarchy process; grey relational analysis

## 1. Introduction

For the past few years, the flourishing of gig economy and crowdsourcing triggered a transformation in the field of product design [1]. More and more companies have realized that customers are important external innovation sources, and expected to establish long-term relationships with them; some of the enterprises have started to collaborate with customers to design/develop new products or services [2]. The advancement of Internet technologies have offered a stable channel for consumers to

participate in design activities. Many firms, such as 3M, P&G, Haier, and Dell, have launched online platforms, i.e., crowdsourcing design communities (CDCs), to encourage customers to contribute contents (e.g., designs, ideas) via posting topics and messages, which is the main pathway for users to take part in enterprises' new product development (NPD) projects [3,4]. For instance, Dell has initiated a crowdsourcing platform, named Ideastorm (URL: https://www.delltechnologies.com/en-us/index.htm), for NPD. Dell posts its needs (e.g., bottleneck problems the company encounter during the process of NPD) on Ideastorm, and invites users to contribute ideas, designs, experience and knowledge to solve these issues. With the Ideastorm strategy, Dell's NPD productivity has increased by almost 50 percent; many of Dell's best-selling products are coming from Ideastorm [5]. The CDC extends the communication channel between companies and customers, which provides a new approach for enterprises to continuously acquire new ideas and external knowledge [6]. Moreover, as an important innovation force, users can exert significant effects on the success of NPD.

In the research field of sustainable crowdsourcing design, scholars have acknowledged that companies should employ the users who have rich experience, extensive knowledge and great skill concerning technologies and usage of various types of products as long-time partners to develop new products [7]. With these individuals' assistance, enterprises can effectively decrease the research and development (R&D) costs, shorten development cycles, increase the probability of project success, and improve the design efficiency [8]. Eric von Hippel defined these individuals as lead users and considered that they are customers of a product/service that current experience needs will be general in a marketplace in the future and who also benefit significantly if they obtain a solution to these needs [9]. These individuals may speed up product R&D, and promote the sustainable development of enterprises. Lead users have two typical features: first, their needs represent the development trends of the product/service; second, they are keen to participate in design projects to translate their needs into new products/services [10]. Since lead users are the most valuable participants in sustainable crowdsourcing design activities, how to identify these individuals with high efficiency is an important research issue in the area of open innovation and information system.

The CDC is the main medium for enterprises to organize crowdsourcing design events. Within the community, the operators may launch product development projects and design challenges. Users can participate in their preferred activities, and contribute content via posting feedbacks and messages to solve the design-related issues. Additionally, the enterprise also encourage users to interact with others in the community; many members often initiate topic discussions on technology, product improvement, usage experience, requirements etc., and reply to other users' posts. The members' contributed content can well reflect their capabilities, expertise and active

degree, which may assist the company managers to discover key users (e.g., lead users, opinion leaders) [11].

Research on discovery of lead users in social media (e.g., social network sites, online communities) is still at initial stages; in particular, strategies to identify these users in the context of online communities have not been sufficiently explored [12,13]. The existing literature primarily introduced two kinds of methods to discover lead users: manual screening and ordering algorithm based on influence rank [11]. The main goal of the former method is to let the community members recommend lead users. It has several tools such as surveying, interviewing and discussing. Although the manual screening method has been widely applied in the research area of product development, it has some limitations, such as low coverage, high cost, and strong subjectivity [9]. The latter method aims to identify the lead users by evaluating the frequency of content contributed by community members and their social influence. Such method can be utilized to handle large samples sizes, but their accuracy is relatively low, and most of the users identified by ordering algorithm are opinion leaders who may have a poor understanding of technologies and usage of products [13]. In order to accurately and efficiently recognize the lead users in the context of the CDC, this work proposes an integrated method combined with user-contributed content and online behavior (mainly contribution behavior) analysis. The content analysis is performed using text-mining technology while the behavior analysis is implemented adopting the statistical tool of user online behavior. The main contributions of this paper are as follows:

(1) We propose an integrated criteria which measure the expertise and active degree of individuals.

(2) Text-mining techniques are applied to extract product-related, innovation-related, user demand-related, etc. information from user-contributed content in CDC.

(3) A ranking system based on fuzzy analytic hierarchy process (FAHP) and grey relational analysis (GRA) is developed to identify lead users.

(4) We demonstrate the efficacy of our proposed methodology utilizing a case study of user behavior data from a well-known CDC in China.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on lead user identification. In addition, Section 3 describes our proposed method in detail. Next, Section 4 presents the case study and discusses the experimental results. Finally, Section 5 draws the conclusions.

## 2. Related Works

### *2.1. Crowdsourcing Design in the Context of Online Platforms*

The companies that desire to perform sustainable crowdsourcing design should motivate users to continuously contribute knowledge, ideas, designs, and innovations. An approach for enterprises to collect these resources is to run a CDC. Such community

often starts out as consumer support platform (e.g., brand community) in which customers exchange information (e.g., usage attentions, tips) of company's products and evolves into a way by that users can put forward suggestions on product improvement and develop extensions [14,15]. The companies may adopt some of the good ideas contributed by the customers, and develop the products according to their needs [16]. Besides, the enterprises often post the issues they meet in the process of NPD, and encourage users to contribute content to solve these problems. Sometimes the users may provide unconventional but effective solutions. Although CDC has been widely used for implementing open innovation, previous studies have paid little attention to systematically explore the users' features and their participation behavior in such context. Therefore, more in-depth research is necessary.

Another method for firms to initiate sustainable crowdsourcing design is to launch an design contest website [16,17]. Within the platform, enterprises put their needs (for technology, products, e-commerce, etc.) on certain sections, and members can find the need which is a match for their innovation capability through the retrieval system [18]. The operators will choose the best designs/ideas as solutions for the needs and pay the members for their contributions [19]. For the purpose of getting the reward, members will try their best to beat the opponents. OpenIDEO hosted by IDEO, HOPE hosted by Haier and Cuusoo hosted by LEGO are typical representatives of design contest website [20].

### *2.2. Manual Method to Discover Lead User*

The research on lead user in the area of product design and innovation is mainly focused on identifying consumers who contribute innovative ideas which are ahead of market preferences and trends [11,21]. In the offline context, manual screening of a great number of potentially relevant customers is the main method to evaluate and identify lead user [22]. Hippel et al. first probed the identification methodology, and proposed screening and pyramiding method to search lead users [9,23]. To perform screening, a representative sample or a predefined population is screened for users who satisfy a certain criterion via questionnaires [9,13]. To discover the real lead users, the examined sample should be sufficiently large. Pyramiding is the improvement of screening; it is a more targeted method that dramatically reduces the effort of screening. To implement pyramiding, researchers must build the pyramid of expertise which contains three layers: the lead users, the users who have good knowledge of product and can find the lead users, and the users who have an understanding of the product domain and may find the advanced experts [23]. The researchers may contact any users of these layers, and follow the chain of user recommendations to find the next level users. This method is effective since respondents who have strong interests in certain topics tend to know the senior experts in the area [12].

Many scholars have further developed Hippel's strategy: Lüthje empirically explored the features of innovation participants, and considered that researchers should incorporate more indicators such as user influence, innovation ability and forward-looking expectations to perform screening [24]. Morrison et al. applied leading-edge status to measure the users' level of expertise; their research results revealed that applications innovativeness is one of the most important features of lead users [25]. Tietz et al. proposed signaling method which utilized advertising tools to discover lead users; such approach can broadcast the survey information which may attract more target users [26]. Hienerth and Lettl explored the measurement of the lead user construct, they considered that social media, data mining, and modern search technologies may be employed to improve the effectiveness of the manual method [27]. These works provide feasible manual methods for scholars to identify lead users in the offline context; however, such methods have some major shortcomings: time-consuming procedure, high search costs, low sampling efficiency, strong subjectivity and cannot contact all the lead users in the user space [11][12][28]. Hence, the manual method should be improved to suit the online environment.

### *2.3. Ordering Algorithm to Identify Lead User*

With the rapid development of information technologies, many scholars considered that monitoring social media may replace the manual method to collect the information of customers [21]. Tang et al. proposed user-rank algorithm which combined content and network analysis to discover influential members in online communities [29]. Song et al. developed influence-rank algorithm to identify opinion leaders in Blogospheres; their method adopted social networks among community members which are not always practicable in some online platforms [30]. Zhao et al. utilized machine learning technique to find lead users in the context of virtual cancer community; such method can identify key users in certain domain, but it is difficult to train the algorithm [31]. Tuarob and Tucker developed a matching algorithm which connected the relationships between lead users and product features; their algorithm can identify the lead users who have special interests in certain area [11]. Pajo et al. proposed a classification model to subdivide the users; such method can well identify the characteristics of different users [12]. To optimize the identification of lead users, scholars have explored the additional features of lead users in the context of online communities. For instance, most lead users are influential and active members in the cyberspace, they often generate product-related and service-related contents, and their opinions represent the vast majority of users' perceptions [32][33]. Although the ordering algorithms suggested by the existing works have improved the efficiency of lead user discovery, they have some drawbacks: (1) the definition of lead users in most of these works relates to how the users' views propagate throughout the online platform, while the lead users in the innovation field should be the individuals who have extensive knowledge and unknown demands; that

is to say, most of the existing methods are developed to identify opinion leaders [11][13]. (2) Most of the approaches need the network connectivity among members which is not always available in communities [11]. Thus, the ordering algorithm should be further developed to analyze the characteristics of lead users from the perspectives of knowledge and demand.

## 3. Methods

### *3.1. Research Framework*

The CDC is constructed based on the online forum, that motivates users to generate contents and interact with other members; within the CDC, enterprises disclose the design-related problems they face through posts, and encourage customers, fans, experts, et al. to contribute (through post image, text and videos in the community) to settle these design challenges [3]. Additionally, operators also encourage members to post their demands in the forum, so that the companies can understand the customers' trends of demand development, and initiate new projects of product development [34]. Hence, the CDC users may generate a large amount of content which can reflect their capabilities.

This work develops an identification method which considers multiple characteristics of lead users, including active degree, expertise and quality of demands. The method contains three steps: first, we develop a criterion system based on previous literature to measure the features of potential lead users; second, text-mining techniques are utilized to collect user-contributed content and user's online behavior statistics from the CDC, and then, we apply the FAHP to evaluate the weight of the indicators and establish the reference sequence of criteria; third, the GRA is employed to calculate the correlation between candidate set (the potential users' indicator set) and reference sequence, and the users are ranked based on their correlations. At last, the top-ranking users (the scales are decided by the enterprises) will be considered as lead users. After these steps, a case study is performed to verify the efficiency of the proposed approach.

### *3.2. The Criterion System for Lead User Identification*

Since CDC users generate design-related content primarily through posting topics and feedbacks, the frequency of content contribution and the correlation between content and innovation may reflect the user's leading-edge-status [2][12][24][25]. Additionally, some researchers suggested that the user's influence in the community may be a key characteristic of lead user [32][33]. However, from the analysis of online innovation platforms (e.g., CDCs, Crowdsourcing websites), we noticed that many professional discussion topics posted by users who have extensive knowledge and great skill on technologies and innovation attract very little attention from other members. These users meet the criteria of lead user identification proposed by Hippel, but their

visibility in the community is relatively low. Thus, we considered that social attributes (e.g., individual's influence) are not the essential features of lead users.

Following previous studies [12][24][25], we employ characteristics of contribution behavior (e.g., contribution frequency) and correlations between user-contributed content and product, innovation, design, and technology as evaluation indicators to measure the individual's leading status. In particular, Guo et al. considered that the ranking system of the online community, which is a kind of statistical tool, can well reflect the features of the user's online behavior [3]. Hence, we apply this system to analyze user's contribution behavior. Besides, text-mining and analysis techniques are utilized to evaluate the relationships between user's contribution and innovation.

### *3.3. The Calculation of Evaluation Indicators*

#### 3.3.1. The indicators of features of user's contribution behavior

Nowadays, most of the online communities have developed the ranking system which can reflect the member's reputation, active degree and community influence by analyzing their online behavior such as posting, replying, likes, sharing and consumption. The system give the user a corresponding rank based on the statistics of individual's behavior, and assist the operator to manage the community. Table 1 shows the introduction of indicators from common ranking system.

**Table 1.** Introduction of indicators from common ranking system.

| Indicators | Introduction |
|---|---|
| Rank | The value of user rank. |
| Title | Virtual honor obtained by users when they reach a certain level. |
| Point | A behavioral credential that users obtain by using community, browsing, posting, purchasing goods, etc. |
| Contribution Value | Reflecting the depth of users' participation in online activities. |
| Virtual Currency | The rewards that users receive through contributing behavior, which can be used for virtual consumption. |

As shown in table 1, contribution value reflects the depth of community members' participation in online events. In the context of open innovation platform, this indicator also reflects the breadth and depth of users' expertise and usage experience. Point is calculated based on statistical information of user's online behavior, which describes the frequency of posting, replying, etc. and reflects the individual's active degree. Rank is a comprehensive reflection of user's contribution level in the community, which

reflects the relative position of the user in the member group. These three indicators can well reflect the user's leading-edge status. Therefore, the contribution value, point and rank are applied to measure the behavior characteristics of users in this work.

3.3.2. The indicators of correlations between user-contributed content and innovation

**Table 2.** The indicators of criterion system for lead user identification.

| **Standard categories** | **Introduction** | **Calculation basis (Indicators)** |
|---|---|---|
| features of user's contribution behavior | These indicators can be employed to measure user's interaction level, contribution frequency, product usage, etc., which reflect the individual's active degree and experience. | contribution value<br>point<br>rank |
| correlations between user-contributed content and innovation | These indicators can be utilized to reflect the user's capabilities of innovation, expertise, hierarchy of needs, usage experience, etc. | words of product features/ attributes<br>words of product issues/ evaluative lexis<br>emotional words<br>effective length of comments<br>timeliness of comments |

The projects of NPD initiated by the enterprise are mainly carried out around the structure, function, and appearance; hence, the online comments from users which contain these contents are often focused by the product developers [35]. Users who contribute such content may likely include lead users. Long explored the relationships between emotional factors and product development, he considered that the users' emotional tendencies may exert significant influence on new product design [36]. Li proposed an ordering algorithm which can be used to evaluate the effectiveness of user comments; he suggested that the individual's reputation, number of thumbs up, timeliness, effective length, words of product features (i.e., attributes) and emotional words may be applied to estimate the correlation between online comments and innovation [37]. Therefore, following prior literature and combining the analysis results of open innovation communities, we utilize words of product features and issues, emotional words, effective length and timeliness of comments as indicators to evaluate the user-contributed content. Table 2 shows the introduction and calculation basis of

indicators of contribution behavior characteristics and correlations between user-contributed content and innovation.

The quantitative methods for calculating the indicators of correlations between user-contributed content and innovation are as follows.

1. The calculation of indicator of words of product features.

Attributes reflect the inherent characteristics of product, such as structure, appearance, etc.; most of these words are nouns. When users post their opinions (such as evaluation, demand, etc.) on products in the community, they often use words of product features to describe them. Therefore, we considered that when the comments contain product attributes, the comments may reflect product-related content (e.g., use experience, product problems, improvement suggestions, etc.), and the more attribute words are included, the more information is transmitted, the greater the auxiliary role for product development and improvement, and the higher the effective contribution level of users.

In this paper, we employ the single comment (i.e., a complete comment) posted by users as the analysis objects. A self-developed spider tool is used to collect users' comments from CDC, and we apply jiebaR to segment the collected Chinese texts into words and tag them with proper Part-of-Speech (PoS) tags (e.g. noun, verb, adverb and adjective). After deleting the stopwords and punctuation characters, the single comment is transformed into a word set $\boldsymbol{V_{1i}}=(v_{11},v_{12},\ldots,v_{1n})$ which contains $N_a$ words. The R programming language is utilized to match the words in $\boldsymbol{V_{1i}}$ one by one with the words of product features in a lexicon $\boldsymbol{V_2}$ which is developed by Institute of Computing Technology, Chinese Academy of Sciences [38]. When a word is matched, the number of attributes of the comment is increased by one. We use $N_b$ to represent the number of attributes in the single comment.

2. The calculation of indicator of words of product issues

The words of product issues (i.e., evaluative lexis) are often used to describe consumers' intuitive perception of the product functions, appearance and other attributes. For instance, "too large" in "car fuel consumption is too large" is the user's intuitive feeling towards car fuel consumption. These words often reflect user's demand expectations and their attitude towards the product. Therefore, we consider that when a comment contains words of product issues, the users may post content about the product use experience, product defects, and personal needs. The more words of product issues that are included, the more detailed the description of the product problem, and the deeper the engagement of users.

The product issues often described with adjectives and verbs, which are usually used with adverbs. For example, within the comment "the motor performance is not good", "motor" and "performance" are words of product features; the adverb "not"

modifies the adjective “good”, and they constitute the word of a product issue. We apply R programming language to identify the adverbs, adjectives and verbs in the comments. When an adjective or a verb appears in $\boldsymbol{V_{1i}}$, the number of words of product issues of the comment is increased by one. Additionally, when an adverb appears in $\boldsymbol{V_{1i}}$ together with an adjective or a verb, the number of words of product issues is also increased by one. We use $N_c$ to represent the number of words of product issues in the single comment.

3. The calculation of indicator of emotional words.

When users post their opinions on products in communities, they often express their emotional tendencies through the vocabulary they used. For instance, “perfect”, “good” and “satisfied” express positive emotions, “bad”, “terrible” and “disappointed” express negative ones. The emotion expressed by the users on the functional attributes of the product can be regarded as an open test result for the product [39]. When users express positive emotions, it indicates that the product has satisfied the users’ expectations in a certain aspect, and there is no need to improve the product at the moment; however, when users express negative emotions, it indicates that certain attributes of the product have not reached their expectations, and the enterprise should improve the product as soon as possible. Both positive and negative emotions can reflect individual’s demand tendency and provide important references for product improvement. Hence, when the comment contains emotional vocabulary, it may reflect the users’ evaluation of the product and their demand tendencies. And the more emotional words are included, the stronger the user’s emotional tendency is reflected.

We use R programming language to match the words in $\boldsymbol{V_{1i}}$ with the emotional words in the lexicon $\boldsymbol{V_3}$ of ICTCLAS [38]. When a word is matched, the number of emotional words of the comment is increased by one. We use $N_d$ to represent the number of emotional words in the single comment.

4. The calculation of indicator of effective length of comments.

In the Chinese language environment, the length of an online comment is usually quantified by the number of Chinese characters included in the comment. However, most online comments contain a large number of meaningless contents, and some of them include numerous characters that have nothing to do with innovation. Thus, we should apply the effective length of the comments to evaluate the leading-edge status of the users. In this work, the ratio of the number of emotional words, product features, and issues in the comment to the number of words in $\boldsymbol{V_{1i}}$ is utilized as the quantized value of the effective length of the comment. Meanwhile, to reduce the deviation caused by the abnormal length (e.g., too long or too short) of the comment, the logarithm is used to weaken the difference of the denominator, as shown in equation (1):

$$R = \frac{\lg(N_b + N_c + N_d)}{\lg N_a}, \quad (1)$$

5. The calculation of indicator of timeliness of comments.

The timeliness of comments refer to the difference value between the time when the user post comment and the time when the comment is fetched by the researchers; the smaller the value, the higher the timeliness [37]. For product innovation, the more time-sensitive comments reflect the newer needs of users, and the less likely they are to be discovered and resolved by competing companies. The more content the user posts in a period, the higher the user's participation, and the higher the user's leading-edge status.

Meng and Ding suggested that the newer the comments, the higher the credibility [40]. Based on their results, we divide the comments into two groups: the comments posted in the last 3 months and the ones posted 3 months ago. The former's value is 2 while the latter is 1.

### *3.4. The Ordering Algorithm of Evaluation Indicators*

The identification of lead users is mainly achieved by ranking the leading-edge status of the users. The rules are as follows: the weight of each evaluation indicator is calculated by FAHP; then the optimal value of each indicator in the research sample is selected to form a reference sequence, and the correlation between each candidate user's indicator sequence and the reference sequence is calculated by GRA. The greater the degree of association, the higher the user's leading-edge status. Based on the research purpose and requirements, a certain relevance threshold can be set to distinguish between lead users and normal users.

#### 3.4.1. The calculation of indicator weight based on FAHP

FAHP is a research method widely applied in the analysis and decision-making of complex systems [41]. It can simplify complex problems into ordered hierarchical structures. In this work, such method is used to determine the weights of various indicators. The analysis steps are as follows:

Step 1 Constructing judgment matrix.

The judgment matrix is a reflection of individuals' thinking and judgment, it can be employed to collect the users' opinions on the weight of the indicators. Since the users of the CDC are the analysis objects of this work, the data that constitutes the judgment matrix mainly comes from the network survey of community members. During the investigation process, the judgment matrix will be sent to the user's mailbox in the form of a web questionnaire. Then, the members compare and score the importance of indicators according to their experience and feelings. The scale applied in the matrix is the 0-0.5-1 standard, its descriptions are shown in Table 3.

**Table 3.** Scale descriptions.

| Scales | Definition | Introduction |
|---|---|---|
| $t_{ij}$=1 | Important | The indicator *i* is more important than the indicator *j* |
| $t_{ij}$=0.5 | Equally important | The indicator *i* and indicator *j* are equally important |
| $t_{ij}$=0 | Unimportant | The indicator *j* is more important than the indicator *i* |

$t_{ij}$ is the judgment value

Step 2 Constructing fuzzy consistent matrix.

After constructing the judgment matrix, we utilize the method proposed by Zhang [42] to transform the matrix into fuzzy consistent matrix.

The rows and columns of the judgment matrix are respectively summed, that is,

$$t_i = \sum_{j=1}^{n} t_{ij}, \qquad i=1,2,\ldots,n \tag{2}$$

$$t_j = \sum_{i=1}^{n} t_{ij}, \qquad j=1,2,\ldots,n \tag{3}$$

After summing, transform each element in the matrix, that is,

$$t'_{ij} = \frac{(t_i - t_j)}{2n} + 0.5, \qquad i, j=1,2,\ldots,n \tag{4}$$

After the transformation, we can get the fuzzy consistent matrix.

Step 3 Calculating indicator weights.

The consistency test is performed to exam the fuzzy consistent matrix, and then, the weight $w_i$ of each indicator $t_i$ is evaluated by the matrix, that is,

$$w_i = \frac{1}{n} - \frac{1}{2\alpha} + \frac{1}{n\alpha}\sum_{j=1}^{n} t'_{ij}, \qquad i, j=1,2,\ldots,n \tag{5}$$

In the equations (2)-(5), n represents the number of indicators. Additionally, in order to improve the resolution of the sorting result, researchers often set α = (n-1)/2. Finally, the average values of each indicator weight $W = (w_1, w_1,\ldots, w_n)$ are obtained.

3.4.2. The ranking of user's leading-edge status based on GRA

GRA is a multi-factor statistical analysis method [43]. Its basic idea is to determine whether the correlation between multiple sequences and the reference sequence is close, and then describe the size, strength and order of the relationship among factors according to the degree of association. For lead users, the greater the value of each indicator, the higher the leading-edge status. Compared with the traditional statistical analysis methods, the advantages of GRA are mainly as follows: GRA is analyzed according to the development trend of the research objects. Therefore, there is no excessive requirement for the size of the sample; no data is required to have a typical distribution law; the calculation amount is relatively small; and the result is in good agreement with the qualitative analysis result [44]. The analysis steps are as follows:

Step 1 The dimensionless processing of the data.

$X_k(i)$ and $Y(i)$ respectively represent the user's indicator sequence and the reference sequence, $k$=1, 2, . . . , $m$, $i$=1, 2, . . . , $n$. $m$ represents the number of users to be ranked, and $n$ represents the number of indicators. Since different indicators often have different dimensions and orders of magnitude, direct comparisons cannot be made and normalization is required. In the related research of GRA, scholars often use the min-max method for dimensionless processing. However, since the data composition in the criterion system is quite complicated, the magnitude of the difference between the different indicators is very large, the min-max method is not applicable in our work. Hence, we employed the averaging method proposed by Song et al. [45] to perform the dimensionless processing, that is,

$$x'_{ki} = \frac{x_{ki}}{\overline{x_i}} \quad k=1,2,\ldots,m, \quad i=1,2,\ldots,n \tag{6}$$

$$y'_i = \frac{y_i}{\overline{x_i}} \quad i=1,2,\ldots,n \tag{7}$$

$X_{ki}$ is the quantized value of the $i$-th indicator of the $k$-th user; $x_i$ is the average value of the $i$-th indicator of the candidate users' sequence; $y_i$ is the quantized value of the $i$-th indicator of the reference sequences. The users whose indicator values are the same as the corresponding indicator values in the reference sequence need to be eliminated to avoid invalid results.

Step 2 The calculation of the grey correlation coefficient.

The formula for calculating the grey correlation coefficient is as shown in equation (8):

$$r_{ki} = \frac{\Delta 1 + \zeta \Delta 2}{\Delta 3 + \zeta \Delta 2} \quad k=1,2,\ldots,m, \quad i=1,2,\ldots,n \tag{8}$$

In the equation, $\Delta 1 = \min_{k \in M} \left\{ \min_{i \in N} \left| Y'(i) - X'_k(i) \right| \right\}$, $\Delta 2 = \min_{k \in M} \left\{ \max_{i \in N} \left| Y'(i) - X'_k(i) \right| \right\}$, $\Delta 3 = \left| Y'(i) - X'_k(i) \right|$, $M = \{1,2,\ldots,m\}$, $N = \{1,2,\ldots,n\}$. $\zeta$ is the resolution coefficient, and its value is taken in the interval [0,1]. Normally, $\zeta$ = 0.5.

Step 3 The calculation of grey correlation.

The weighted method is used to calculate the gray correlation, and the formula is as shown in equation (9):

$$d_k = \sum_{i=1}^{n} w_i r_{ki} \qquad k \in M \tag{9}$$

In the equation, $w_i$ is the value of indicator weight obtained by FAHP, $\sum_{i=1}^{n} w_i = 1$. Sorting the correlation of each user, and then, we can get the ranking of the user's leading-edge status. Setting the relevance threshold (dynamic) according to the semantic environment and specific purpose of the research. Then, we can distinguish between lead users and normal users.

## 4. Empirical Study

### *4.1. Data Crawling*

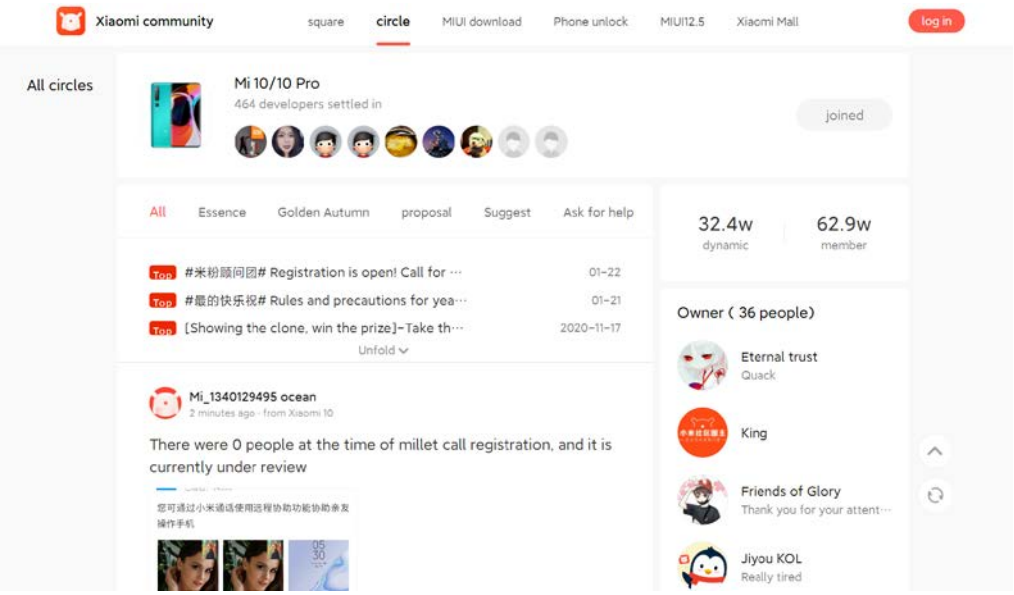



Figure 1. User interface of Xiaomi Forum (Section of MI10/10 Pro).

In this work, a case study of lead user identification is conducted to verify the effectiveness and practicality of our proposed method. We examine the method with samples collected from a CDC, Xiaomi Forum, which has over 50 million registered members, and about 65 percent of them are active users [3]. Xiaomi is a mobile Internet company focused on designing and manufacturing smartphones; its CDC collects a great many valuable ideas and designs from the community members [2]. Hence, it is an ideal source for the research. Figure 1 shows the user interface of Xiaomi Forum.

We applied a self-developed spider program to collect 9500 users' comments (including topic posts and feedbacks) and their recent ID information (including contribution value, rank and point) from 2018/11/1 to 2019/5/20. The R programming language and SPSS tool were utilized to perform the analysis.

*4.2. Data Analysis and Results*

Among the related comments published by the 9500 users, the maximum number of occurrences of words of product features in the user comments is 1722, the number of words of product issues is 207, and the number of emotional words is 1079. In terms of effective length of comments, the statistical value is processed by R programming language, and the maximum quantized value is 0.697. In terms of timeliness of comments, the maximum average value of all users' posts is 1.7. In terms of contribution value, the highest value is 11328. In terms of rank, the highest value is 8. In terms of point, the highest value is 213176. We applied these values as the reference sequence, and calculate the correlations of all the users. Then, we sorted the users based on their correlations. Since Hippel considered that only about 3 percent of customers are lead users [9], we took the top 3 percent of the 9500 users as the lead user.

In order to verify the validity of the method proposed in this study, we compared the recognition results of the manual method with our approach. Following the method suggested by Brem and Bilgram [13], we selected 500 Xiaomi Forum users who have used the community for more than one year, and sent a survey invitation to them through mail system. We asked them to select the 50 most representative community members based on their feelings and experience. The respondents rated these users in terms of expertise, experience, demand and participation level (each scored 1/4), and the top 20 users are ranked as the lead users of the community. A total of 272 respondents have returned valid information. In order to ensure the effectiveness of the comparison verification, a 45-day online behavior tracking was conducted for these 20 lead users. Through the analysis, we found that each of these users posts at least 10 topics per week, and most of the posts are related to market, product and technology. These posts were analyzed by 7 scholars and technical experts from enterprises/colleges, they found that in addition to the 5 users ranked 1st, 4th, 10th, 13th and 19th (these 5 users post a large amount of topics, and they are well-known in the community, but

only a small part of their posts are related to technology, market, experience, demand, etc. Therefore, they are just simple active users.), the other 15 users had a high leading-edge status. As shown in Table 4, the top 10 lead users identified by our method are compared with the lead users recognized by the manual method. The ID in the table is the user's identity number in Xiaomi Forum.

**Table 4.** Results comparison of the two methods.

| User ID | The users' correlations calculated by the method proposed in this work (From high to low) | The user ranking analyzed by the manual method |
|---|---|---|
| 139359812 | 0.792 | 3 |
| 1069696768 | 0.756 | 6 |
| 179526422 | 0.707 | 5 |
| 95572881 | 0.633 | 11 |
| 437500596 | 0.622 | / |
| 158179452 | 0.609 | 8 |
| 103017361 | 0.574 | 12 |
| 139172452 | 0.518 | 7 |
| 23957255 | 0.462 | / |
| 148869817 | 0.430 | 9 |

/ represents the users who are not in the top 20

The results revealed that the identification method proposed in this study has good precision and recall rate. Besides, the comparison showed that manual method is highly dependent on human resources: in the early stage, it takes a lot of time to find qualified respondents; in the analysis stage, it needs to carry out cumbersome data processing to reduce the adverse impact of the user's subjective feelings on the survey results. The whole process is complicated, and it is difficult to guarantee the quality of the survey. Compared with the manual method, the approach suggested by our work mainly uses the contribution content and statistical information retained by the community members to identify the lead users. It can automatically process large-scale data and has no strict requirements for the respondents. Hence, our method has advantages in terms of efficiency and accuracy.

## 5. Conclusions and Limitations

Lead users are the most valuable customer groups in the NPD. Therefore, accurately identifying and locating lead users is of great significance for enterprises to effectively organize and manage sustainable open design activities. This study proposes

a lead user identification method based on user behavior data and contribution content analysis and constructs a criterion system to evaluate the user's leading-edge status. The effectiveness of the proposed method in this work is verified by comparative analysis. Compared with the manual method, our approach has several advantages, such as high efficiency, accuracy, and coverage.

This study is restricted by some limitations. The method in the paper is to identify the lead users by sorting the correlations of the community members. The greater the degree of relevance, the higher the user's leading-edge status. However, how to select the relevance threshold to distinguish between lead users and normal users is not yet clear. Based on Liao's research results [46], we considered that the operators may determine the threshold by two methods: 1. The total amount (M) of valid information in the content contributed by the user can be used as the threshold value. When the user's contributed valid content is more than M, she/he can be regarded as the lead user. The selection of the M value depends on the improvement rate of the product proposed by the enterprise. 2. According to the ranking of correlations, the top P percent of the candidate users are lead users. The selection of the P value depends on factors such as the size of the research sample, the extent to which the company intends to improve the product, and the size of the potential customers who may purchase the improved product. Future research may verify these two threshold determination methods and discuss the context in which each method applies.

## References


1. Howe J. The rise of crowdsourcing. *Wired Mag*. 2006, *14*, 1-4.
2. Liang, R.Y.; Guo, W.; Zhang, L. H.; Wang, L. Investigating Sustained Participation in Open Design Community in China: The Antecedents of User Loyalty. *Sustainability*. 2019, *11*, 2420-2439.
3. Guo, W.; Liang, R.Y.; Wang, L.; Peng, W. Exploring sustained participation in firm-hosted communities in China: The effects of social capital and active degree. *Behav. Inf. Technol*. 2017, *36*, 223–242.
4. Liang, R.Y.; Zhang, L.H.; Guo, W. Investigating active users' sustained participation in brand communities: Effects of social capital. *Kybernetes*. 2019, *48*, 2353-2372.
5. Bayus, B.L. Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. *Management science*. 2013, *59*, 226-244.
6. Kyriakou, H.; Nickerson, J.V.; Sabnis, G. Knowledge reuse for customization: metamodels in an open design community for 3d printing. *Social Science Electronic Publishing*, 2017, *41*, 315-332.
7. Fernandes, S.; M. Cesario.; J. M. Barata. Ways to open innovation: Main agents and sources in the Portuguese case. *Technology in Society*, 2017, *51*, 153-162.
8. Lilien, G.L.; Morrison, P.D.; Searls, K.; Sonnack, M.; Hippel, E.V. Performance assessment of the lead user idea-generation process for new product development. *Management science*, 2002, *48*, 1042-1059.

9. Urban, G.L.; Von Hippel, E. Lead user analyses for the development of new industrial products. *Management science*, 1988, *34*, 569-582.
10. Herstatt, C.; Von Hippel, E. From experience: Developing new product concepts via the lead user method: A case study in a “low-tech” field. *Journal of product innovation management*, 1992, *9*, 213-221.
11. Tuarob, S.; Tucker, C.S. Automated discovery of lead users and latent product features by mining large scale social media networks. *Journal of Mechanical Design*, 2015, *137*, 071402.
12. Pajo, S.; Vandevenne, D.; Duflou, J.R. Automated feature extraction from social media for systematic lead user identification. *Technology Analysis & Strategic Management*, 2017, *29*, 642-654.
13. Brem, A.; Bilgram, V. The search for innovative partners in co-creation: Identifying lead users in social media through netnography and crowdsourcing. *Journal of Engineering and Technology Management*, 2015, *37*, 40-51.
14. Raasch, C., Herstatt, C.; Balka, K. On the open design of tangible goods. *R&D Management*. 2009. *39*, 382–393.
15. Wiertz, C.; de Ruyter, K. Beyond the call of duty: Why customers contribute to firm-hosted commercial online communities. *Organization Studies*. 2007. *28*, 347–376.
16. Jeppesen, L.B.; Frederiksen, L. Why do users contribute to firm-hosted user communities?: The case of computer-controlled music instruments. *Organ. Sci*. 2006, *17*, 45–63.
17. Seidel, V.P.; Langner, B. Using an online community for vehicle design: Project variety and motivations to participate. *Ind. Corp. Chang*. 2015, *24*, 635–653.
18. Jeppesen, L. B.; Lakhani, K. R. Marginality and problem-solving effectiveness in broadcast search. *Organization Science*. 2010. *21*, 1016–1033.
19. Afuah, A.; Tucci, C. L. Crowdsourcing as a solution to distant search. *Academy of Management Review*. 2012. *37*, 355–375.
20. Li, M.; Jia, S.; Du, W. Fans as a source of extended innovation capabilities: a case study of xiaomi technology. *International Journal of Information Management*. 2019. *44*, 204-208.
21. Bilgram, V.; Brem, A.; Voigt, K. User-centric innovations in new product development — systematic identification of lead users harnessing interactive and collaborative online-tools. *International Journal of Innovation Management*, 2008, *12*, 419-458.
22. Belz, F.M.; Baumbach, W. Netnography as a method of lead user identification. *Creativity & Innovation Management*, 2010, *19*, 304-313.
23. von Hippel, E.; Franke, N.; Prügl, R. Pyramiding: efficient search for rare subjects. *Res. Policy*, 2009, *38*, 1397–1406.
24. Lüthje C. Characteristics of innovating users in a consumer goods field: An empirical study of sport-related product consumers. *Technovation*, 2004, *24*, 683-695.
25. Morrison P.D.; Roberts J.H.; Midgley D.F. The nature of lead users and measurement of leading edge status. *Research policy*, 2004, *33*, 351-362.
26. Tietz, R.; Füller, J.; Herstatt, C. Signaling: an innovative approach to identify lead users in online communities. International Mass Customization Meeting 2006, Hamburg, Germany, 2006.
27. Hienerth, C.; Lettl, C. Understanding the Nature and Measurement of the Lead User Construct. *Journal of Product Innovation Management*, 2017, *34*, 3-12.
28. von Hippel, E.; Thomke, S.; Sonnack, M. Creating Breakthroughs at 3M. *Harv. Bus. Rev*.,1999, *7*, 47–57.

29. Tang, X.; Yang, C. Identifying Influential Users in an Online Healthcare Social Network. 2010 IEEE International Conference on Intelligence and Security Informatics (ISI), Vancouver, Canada, 2010.
30. Song, X.; Chi, Y.; Hino, K.; Tseng, B. Identifying Opinion Leaders in the Blogosphere. Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM'07), New York, U.S.A., 2007.
31. Zhao, K.; Qiu, B.; Caragea, C.; Wu, D.; Mitra, P.; Yen, J.; Greer, G.E.; Portier, K. Identifying Leaders in an Online Cancer Survivor Community. 21st Annual Workshop on Information Technologies and Systems (WITS'11), Shanghai, China, 2011.
32. Schreier, M.; Oberhauser, S.; Prügl, R. Lead users and the adoption and diffusion of new products: insights from two extreme sports communities. *Mark. Lett*. 2007, *18*, 15–30.
33. Schreier, M.; Prügl, R. Extending lead-user theory: antecedents and consequences of consumers' lead userness. J. Prod. *Innov. Manag*, 2008, *25*, 331–346.
34. Brem A.; Bilgram V. The Search for Innovative Partners in Co-Creation: Identifying Lead Users in Social Media through Netnography and Crowdsourcing. *Journal of Engineering & Technology Management*, 2015, *37*, 40-51.
35. Liang, R.Y.; Guo, W.; Yang D.Q. Mining product problems from online feedback of Chinese users. *Kybernetes*, 2017, *46*, 572-586.
36. Long, S.J. Discussion on design idea based on user demand and product hierarchy. *Journal of Machine Design*, 2013, *30*, 126-128.
37. Li, Z.Y. Study on the Reviews Effectiveness Sequencing Model of Online Products. *Data Analysis and Knowledge Discovery*, 2013, *2013*, 62-68.
38. NLPIR-ICTCLAS. Available online: http://ictclas.nlpir.org/ (accessed on 23 March 2020).
39. Liang, R.Y.; Guo, W.; Zhang, L.H. Exploring oppositional loyalty and satisfaction in firm-hosted communities in China. *Internet Research*. 2020, *30*, 487-510.
40. Meng, M.R.; Ding S.C. Research on the Credibility of Online Chinese Product Reviews. *Data Analysis and Knowledge Discovery*, 2013, *2013*, 60-66.
41. Kubler, S.; Robert, J.; Derigent, W.; Voisin, A.; Le Traon, Y. A state-of the-art survey & testbed of fuzzy AHP (FAHP) applications. *Expert Systems with Applications*, 2016, *65*, 398-422.
42. Zhang, J.J. Fuzzy analytical hierarchy process. *Fuzzy System and Mathematics*, 2000, *14*, 80-88.
43. Deng, J.L. Introduction to grey system theory. *The Journal of grey system*, 1989, *1*: 1-24.
44. Paramasivam, B. Investigation on the effects of damping over the temperature distribution on internal turning bar using Infrared fusion thermal imager analysis via SmartView software. *Measurement*, 2020, *162*.
45. Song, M.S.; Huang, J.; Zhang, S.P.; Qi, B.F. The Research on the Dimensionless Criterion and Methods about the Design of Multi-index Orthogonal Experiment, *Industrial Engineering and Management*, 2014, *19*, 41-46.
46. Liao, X. Liao X.; Li, Z.; Xi, Y.X. The Modeling and Analyzing Methods of Weighted Knowledge Network for Domain Knowledge Based on Keywords Clustering. *The Open Cybernetics & Systemics Journal*, 2014, *8*, 990-997.